TREE BASED METHOD FOR CENSORED DATA

Anaranya Basu

Department of Statistics, St.Xavier's University

1. INTRODUCTION TO CENSORED DATA
2. TREE BASED METHOD IN BRIEF
3. Theoretical Part and Splitting for censored Data
4. OPTIMAL PRUNING
5. ACCURACY CHECKING

1. INTRODUCTION TO CENSORED DATA
2. TREE BASED METHOD IN BRIEF
3. Theoretical Part and Splitting for censored Data
4. OPTIMAL PRUNING
5. ACCURACY CHECKING

INSIGHTS OF SURVIVAL ANALYSIS

- Survival analysis is a set of statistical models and methods used for estimating time until the occurance of an event (or the probability that an event has not occured).
- Statistical methods need data, but complete data may not be available, i.e. the exact time of the event may be unknown for certain reasons (the event did not occur before the end of the study or it is unknown whether it occured). In this case, lifetimes are called censored.
- The data are censored from below (left censored) when below a given value the exact values of observations is unknown. Right censored data (censored from above) does not have exact observations above a given value. Further in this paper, right censoring is considered.

INSIGHTS OF SURVIVAL ANALYSIS

- The problems studied with the help of survival analysis are formulated in terms of survival function (that is complementary distribution function)

$$S(t) = P(T > t)$$

- The distribution of T may also be characterized with so called hazard function,

$$h(t) = -\frac{\partial}{\partial t} \log s(t)$$

INSIGHTS OF SURVIVAL ANALYSIS

- A parametric model assumes a distribution function, and its parameters are estimated based on the available data. Also we may find empirical distribution function and then use its complement as the survival function. Non parametric methods called the Kaplan-Meier estimator (Kaplan and Meier, 1958) and Nelson-Alan (Nelson, 1972) estimator are more powerful. The Kaplan-Meier estimator has the form,

$$S(t) = \prod_{i:t_i<t}(1 - \frac{d_i}{n_i})$$

where t_i is time of the event, d_i is number of events that occurred at time t_i, and n_i number of events after time t_i (or unknown at t_i). Nelson-Aalen esimator applies the same idea to the cumulative hazard function $H(t) = \int_0^t h(s)ds$ and then transforms it to the estimation of the survival function.

1. INTRODUCTION TO CENSORED DATA
2. TREE BASED METHOD IN BRIEF
3. Theoretical Part and Splitting for censored Data
4. OPTIMAL PRUNING
5. ACCURACY CHECKING

REGRESSION TREE

- Tree based approach is a well known supervised method which actually classifies data and partition a data set into smaller groups and then fit a simple model (constant) for each subgroup.
- In case of censored data where the data comes in a incomplete format tree based method is helpful sometimes to know the survival probability with some important risk factors.
- Regression trees are highly interpreretable, which can be important for understanding the factors that influence survival or time-to-event outcomes in censored data.

NOTATION AND DATA SPLIT UP

- An observation will be distributed as the vector (T,δ,X)
- where T is the time under observation, δ is an indicator of failure, and X is a vector of M covariates.
- Let U denote the true survival time having cumulative distribution function F.
- V be the true censoring time with cumulative distribution function C. Then assume $\delta = I_{U \leq V}$, where $I_.$ is the indicator function of the set .

NOTATION AND DATA SPLIT UP

- Define the observed time T = min(U, V).
- Assume also that U and V are independent given X, for identifiability reasons.The learning sample consists of the set of iid vectors :

$$(T_i, \delta_i, X_i) : i = 1, 2, \ldots, N$$

SURV() FUNCTION IN R

This whole process is done in R by the **Surv()** function.

- Surv(time, status)- right censored data
- Surv(time, endpoint==death) - right censored data, where the status variable is a character or factor.
- Surv(t1, t2, status) - counting process data
- Surv(t1, ind, type= left) - left censoring

1. INTRODUCTION TO CENSORED DATA
2. TREE BASED METHOD IN BRIEF
3. Theoretical Part and Splitting for censored Data
4. OPTIMAL PRUNING
5. ACCURACY CHECKING

LOGRANK STATISTICS

- The algorithm splits the data into groups with differing survival. Because the logrank test has been used extensively in the analysis of censored survival data, it is a logical choice for measuring dissimilarity in survival between two groups.
- Let the number of patients in the two groups be n_1, and n_2. Let $Y_1(u)$ and $Y_2(u)$ be the number of individuals at risk in each group at time u

LOGRANK STATISTICS

- $Y_j(u)= \sum_{K \in R_j} \mathrm{I}_{T \leq u}$, j=1,2 where R_j is the set of observation labels corresponding to group j.
- Let $\hat{\Lambda}_1(u)$ and $\hat{\Lambda}_2(u)$ be the Nelson Cumulative Hazard estimator for each group.

LOGRANK STATISTIC

- The numerator of the log rank statistic can be expressed as a weighted difference between estimated hazard functions,

$$G = \int_0^\infty [w(u)\frac{Y_1(u)Y_2(u)}{Y_1(u)+Y_2(u)}](d\hat{\Lambda}_1(u)\text{-}d\hat{\Lambda}_2(u))$$

- where $w(.) = 1$.

LOGRANK STATISTIC

- Stability of a statistic under censoring is an important requirement for splitting
- If the statistic becomes more variable in censored data, it will tend to split in regions of heavier censoring
- the statistic can also be represented as the linear function :

$$G(s) = \sum_i^n I_{X_{jt} \leq s}(\delta_i - \hat{\Lambda}_0(t_i))$$

LOGRANK STATISTIC

- where t_i is time under observation and X_{j_i} is the value for covariate j for individual i and s is the split point
- This form of the statistic leads to an efficient updating formula for two split points s_1 and s_2 as,

$$G(s_1) - G(s_2) = \sum_i^n I_{s_1 < X_{j_t} \leq s_2}(\delta_i - \hat{\Lambda}_0(t_i))$$

1. INTRODUCTION TO CENSORED DATA
2. TREE BASED METHOD IN BRIEF
3. Theoretical Part and Splitting for censored Data
4. OPTIMAL PRUNING
5. ACCURACY CHECKING

COST COMPLEXITY PRUNING

- We define a sub tree $T \subseteq T_0$ to be any tree that can be obtained by pruning T_0
- We index terminal nodes by m, with node m representing region R_m.Letting,

$$N_m = \#[x_i \in R_m],$$
$$\hat{c_m} = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i)$$
$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c_m})^2$$

COST COMPLEXITY PRUNING

- we define the cost complexity criterion as ,

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|$$

- The main ideas is to find, for each α, the sub tree $T_\alpha \subseteq T_0$ to minimize $C_\alpha(T)$
- The tuning parameter $\alpha > 0$ governs the trade-off between tree size and its goodness of fit to the data
- Large values of α result in smaller trees T_α

1. INTRODUCTION TO CENSORED DATA
2. TREE BASED METHOD IN BRIEF
3. Theoretical Part and Splitting for censored Data
4. OPTIMAL PRUNING
5. ACCURACY CHECKING

CONCORDANCE INDEX

- The Concordance Index (Harrell Jr et al., 1996) is widely used in survival analysis.
- it measures the fraction of concordant or correctly ordered pairs of samples among all available pairs in the data set.
- The highest value of the metric is one (if the order is perfect), and the value of 0.5 means that the model produces completely random predictions.

CONCORDANCE INDEX

- Concordance index:

$$CI = \frac{\sum_{i,j} 1_{T_j \leq T_i} 1_{\eta_j \leq \eta_i}}{\sum_{i,j} 1_{T_j \leq T_i}}$$

- where T_j is the true time of the event, and η_j is the time predicted by the model.

BRIER SCORE

- This metric is based only on the predicted time of the event
- It does not allow estimating the survival function
- The value of CI does not change when the survival function is biased
- although the predicted time is highly distorted compared to the true time.

BRIER SCORE

- To eliminate this problem, we use a metric called Integrated Brier Score
- Which is defined as -

$$BS(t) = \frac{1}{N}\sum_i (0 - S(t, x_i))^2 \text{ if } T_i \leq t$$
$$BS(t) = \frac{1}{N}\sum_i (1 - S(t, x_i))^2 \text{ if } T_i > t$$

- where $S(t; x_i)$ is the prediction of the survival function at time t for observation x_i with event time T_i.

BRIER SCORE

- To aggregate the BS estimates over all time moments, the Integrated Brier Score is used:

$$IBS = \frac{1}{t_{max}} \int_0^{t_{max}} BS(t)dt$$